INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Comparing Protein 3D Structures Using A_purva

N. Malod-Dognin — N. Yanev — R. Andonov

N° 7464

Novembre 2010

Computational Biology and Bioinformatics

ISSN 0249-6399 ISRN INRIA/RR--7464--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
SOPHIA ANTIPOLIS - MÉDITERRANÉE

Comparing Protein 3D Structures Using A_purva

N. Malod-Dognin*, N. Yanev†, R. Andonov‡

Theme : Computational Biology and Bioinformatics
Computational Sciences for Biology, Medicine and the Environment
Équipes-Projets ABS

Rapport de recherche n° 7464 — Novembre 2010 — 6 pages

Abstract: Structural similarity between proteins provides significant insights about their functions. Maximum Contact Map Overlap maximization (CMO) received sustained attention during the past decade and can be considered today as a credible protein structure measure. We present here A_purva, an exact CMO solver that is both efficient (notably faster than the previous exact algorithms), and reliable (providing accurate upper and lower bounds of the solution). These properties make it applicable for large-scale protein comparison and classification.
Availability: http://apurva.genouest.org
Contact: support@genouest.org
Supplementary information:A_purva's user manual, as well as many examples of protein contact maps can be found on A_purva's web-page.

Key-words: Protein structure comparison, software, combinatorial optimization, integer programming

* noel.Malod-dognin@inria.fr : INRIA Sophia Antipolis - Méditerranée, France
† choby@math.bas.bg : University of Sofia, and IMI at Bulgarian Academy of Sciences, Bulgaria
‡ randonov@irisa.fr : INRIA Rennes - Bretagne Atlantique, and IRISA/University of Rennes 1, France

Centre de recherche INRIA Sophia Antipolis – Méditerranée
2004, route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex
Téléphone : +33 4 92 38 77 77 — Télécopie : +33 4 92 38 77 65

Comparer des structures de protéines avec A_purva

Résumé : La similarité structurale entre protéines donne des renseignements importants sur leurs fonctions. La maximisation du recouvrement de cartes de contacts (CMO) a reçu une attention soutenue ces dix dernières années, et est maintenant considérée comme une mesure de similarité crédible. Nous présentons içi A_purva, un solveur de CMO exacte qui est à la fois efficace (plus rapide que les autres algorithmes exactes) et fiable (fournit des bornes supérieures et inférieures précises de la solution). Ces propriétés le rendent applicable pour des comparaisons et des classifications de protéines à grandes échelles.
Disponibilité : http://apurva.genouest.org
Contact : support@genouest.org
Informations supplémentaires : Le manuel utilisateur d'A_purva, ainsi que de nombreux exemples de cartes de contacts de protéines sont disponibles sur le site web d'A_purva.

Mots-clés : Comparaison de structures de protéines, logiciel, optimisation combinatoire, programmation en nombres entiers

1 Introduction

It is commonly accepted in structural biology that the bigger the similarity between proteins, the higher the probability that they share a common ancestor and possess similar functions. Since the 3D structures of proteins are often more conserved than their sequences, the structural comparison provides significant insights about the protein functions. We focus here on the Maximum Contact Map Overlap (CMO)[5, 4]–a protein structure similarity measure that is robust, captures well the intuitive notion of similarity, and is translation and rotation-invariant. The CMO problem consists in finding a one-to-one correspondence between subsets of residues in two proteins that maximizes the overlap of their contacts. CMO is an NP-hard problem [6] that has been extensively studied by the bioinformatics and computer science communities. The known algorithms can be mainly classified as exact–but with modest performance (LAGR[3], Clique[12] and CMOS[13]), or as heuristics–faster but without guaranty for the precision of the results (SADP[7], MSVNS[11] and EIG7[8]).

A_purva is an exact CMO solver that is both efficient (significantly faster than the other exact algorithms) and reliable (estimates the precision of approximation when the exact solution is not found).

1.1 Contact map overlap maximization

In the folded state of a protein, amino-acids that are distant in the sequence may come into proximity in 3D space and form contacts. This proximity relation is captured by the contact map graph: it is a simple graph $CM = (V, E)$ where the vertices in V correspond to the amino-acids of the protein, and where a contact edge (i, j) in E connects two vertices i and j if and only if the euclidean distance between the corresponding amino-acids i and j in the protein fold is smaller than a given threshold (see figure 1).

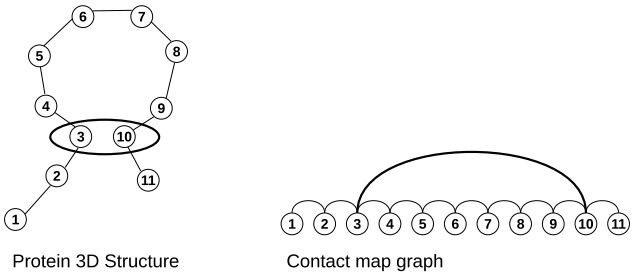

Figure 1: From protein structure to contact map. **Left:** amino-acids 3 and 10 are remote in the protein sequence, but are close in 3D space (the euclidean distance is smaller than a given threshold). We say then that they are in contact. **Right:** vertices 3 and 10 are connected by an edge in the corresponding contact map graph.

In the CMO approach, the similarity between two proteins is given by their Number of Common interatomic Contacts (NCC), as determined by the maximum overlap of their contact maps graphs. The formal definition is as follows. Given two contact maps $CM_1 = (V_1, E_1)$ and $CM_2 = (V_2, E_2)$, let

$I = (i_1, i_2, \ldots i_s)$, $i_1 < i_2 < \ldots i_s$ and $J = (j_1, j_2, \ldots j_s)$, $j_1 < j_2 < \ldots j_s$ be arbitrary subsets of vertices from the first and second contact maps, respectively. Under the matching (also called alignment) $i_k \longleftrightarrow j_k, k = 1, 2, \ldots s$, the edge (k, l) is *common* (an overlap occurs) if and only if both edges (i_k, i_l) and (j_k, j_l) exist. The CMO problem is to find the optimal I and J, where optimality means maximum number of common edges, as illustrated in figure 2.

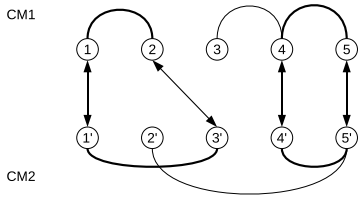

Figure 2: The arrows between the contact maps CM1 and CM2 depict the matching $(1 \leftrightarrow 1', 2 \leftrightarrow 3', 4 \leftrightarrow 4', 5 \leftrightarrow 5')$. This matching activates 2 common contacts–which the maximum for this instance (NCC=2).

2 Using A_purva

2.1 Rationale

A_purva is based on an integer programming formulation of CMO, and it converges to the optimal solution using a branch and bound strategy. At each node, A_purva borders NCC using two numbers derived from a Lagrangian relaxation of the integer programme [1, 2, 9] : a lowerbound LB, which is the biggest number of common contacts found so far in a feasible solution, and an upperbound UB, which is the biggest number of common contacts found in the relaxed problem. The similarity score returned by A_purva is:

$$\mathrm{SIM}(P_1, P_2) = \frac{2 \times LB}{|E_1| + |E_2|} \tag{1}$$

When an instance is optimally solved, the relation $LB = NCC = UB$ holds. Otherwise, $UB > LB$ and the so called relative gap value $RG = (UB - LB)/UB$ gives the precision of the results. This property is very useful in the context of large-scale database comparisons where the execution time is usually bounded.

2.2 Achievements

The results presented here were obtained using two data sets: the Skolnick set (40 protein chains, 5 SCOP families) and the Proteus_300 set (300 protein chains, 10 SCOP families).

Running times. In [1], with the same time limit, A_purva optimally solves 78.2% of the comparison instances from the Skolnick set, while LAGR and CMOS are limited to 20.6%. In average, A_purva is about 1200 times faster than LAGR.

Exactness. The results presented in [8] show that when using the same distance threshold of 7.5Å, the number of common contacts found by MSVNS and EIG7 are smaller than the ones found by A_purva (0.921% and 0.882% of A_purva's values). Using this values in the similarity function (1), A_purva perfectly identifies the SCOP families of the all the 300 protein chains from the Proteus_300 set, while MSVNS makes 3 mistakes and EIG7 makes 6 mistakes (EIG7 correctly identifies the families of all the chains when using distance thresholds $\geq$ 10 Å). These results indicate that the exactness of the solutions affects the quality of the similarity measure.

Family identification. A_purva recently participated in the SHREC'10 contest[10] for identifying the families of 50 "unknown" protein structures within 100 CATH superfamilies–each represented by 10 protein structures. A_purva achieves the highest success rate (88% of correctly classified proteins during the competition), as well as the highest sensitivity and specificity. After the competition, with the exact α-carbon coordinates, the success rate of A_purva is 92%.

Automatic classification. In [1], using the similarity scores returned by A_purva, we automatically obtain the SCOP family classification for the Skolnick set, and a classification that is very similar to the SCOP family one for the Proteus_300 set.

2.3 Inputs and outputs

The web URL – http://apurva.genouest.org – allows to run comparisons, and returns the results by email.

Inputs. A_purva needs either two contact map files, or two pdb files and a distance threshold for defining contacts.

Outputs. A_purva has three levels of output : the first one displays the similarity score only; the second also gives the lower and upper bounds; while the last one displays the optimal alignment as well. If the inputs are pdb files, A_purva can also generate visualisation files for VMD highlighting the matched regions.

Acknowledgement

N. Yanev is supported by projects DO 02359/2008 and DTK02/71. R. Andonov is supported by BioWIC ANR-08-SEGI-005 project. All computations were done on the Ouest-genopole bioinformatics platform (http://genouest.org).

References

[1] R. Andonov, N. Malod-Dognin, and N. Yanev. Maximum contact map overlap revisited. *J. Comput. Biol.*, 18(1):1–15, 2011.

[2] R. Andonov, N. Yanev, and N. Malod-Dognin. An efficient Lagrangian relaxation for the contact map overlap problem. In *WABI '08: Proc. of the 8th Int. Workshop on Algorithms in Bioinformatics*, pp. 162–173. Springer-Verlag, 2008.

[3] A. Caprara, R. Carr, S. Israil, G. Lancia, and B. Walenz. 1001 optimal PDB structure alignments: Integer programming methods for finding the maximum contact map overlap. *J. Comput. Biol.*, 11(1):27–52, 2004.

[4] A. Godzik. The structural alignment between two proteins: Is there a unique answer? *Protein Science*, 5(7):1325–1338, 1996.

[5] A. Godzik and J. Skolnick. Flexible algorithm for direct multiple alignment of protein structures and seequences. *CABIOS*, 10:587–596, 1994.

[6] D. Goldman, S. Istrail, and C. Papadimitriou. Algorithmic aspects of protein structure similarity. In *FOCS '99: Proc. of the 40th Annual Symposium on Foundations of Computer Science*, pp. 512–521. IEEE Computer Society, 1999.

[7] B.J. Jain and M. Lappe. Joining softassign and dynamic programming for the contact map overlap problem. In Sepp Hochreiter and Roland Wagner, editors, *BIRD*, volume 4414 of *LNCS*, pp. 410–423. Springer, 2007.

[8] Pietro Di Lena, Piero Fariselli, Luciano Margara, Marco Vassura, and Rita Casadio. Fast overlapping of protein contact maps by alignment of eigenvectors. *Bioinformatics*, 26(18):2250–2258, 2010.

[9] N. Malod-Dognin. *Protein Structure Comparison: From Contact Map Overlap Maximisation to Distance-based Alignment Search Tool.* PhD thesis, University of Rennes 1, 2010.

[10] L. Mavridis, V. Venkatraman, D. W. Ritchie, N. Morikawa, R. Andonov, A. Cornu, N. Malod-Dognin, J. Nicolas, M. Temerinac-Ott, M. Reisert, H. Burkhardt, and A. Axenopoulos. Shrec-10 track: Protein models. In I. Pratikakis, M. Spagnuolo, T. Theoharis, and R. Veltkamp, editors, *3DOR: Eurographics Workshop on 3D Object Retrieval*, pp. 117–124, Norrköping, Sweden, 2010. The Eurographics Association.

[11] D. Pelta, J. Gonzales, and M. Vega. A simple and fast heuristic for protein structure comparison. *BMC Bioinformatics*, 9(1):161, 2008.

[12] D.M. Strickland, E. Barnes, and J.S. Sokol. Optimal protein structure alignment using maximum cliques. *Oper. Res.*, 53(3):389–402, 2005.

[13] W. Xie and N. Sahinidis. A reduction-based exact algorithm for the contact map overlap problem. *J. Comput. Biol.*, 14(5):637–654, 2007.

Centre de recherche INRIA Sophia Antipolis – Méditerranée
2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Grenoble – Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Saclay – Île-de-France : Parc Orsay Université - ZAC des Vignes : 4, rue Jacques Monod - 91893 Orsay Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399